Journal of Applied Statistics
Vol. 00, No. 00, Month 200x, 1–8

arXiv:1302.0817v1 [stat.ME] 4 Feb 2013

SHORT COMMUNICATION

A statistical testing framework for evaluating the quality of measurement processes

Edgard Nyssen[a*] and Wolfgang Jacquet[b]

[a] *VUB, ETRO Dept., Pleinlaan 2, B-1050 Brussel, Belgium*;
[b] *VUB, EDWE Dept., Pleinlaan 2, B-1050 Brussel, Belgium*

(*v0.1 received October 2012*)

In this paper in which we address the evaluation of measurement process quality, we mainly focus on the evaluation procedure, as far as it is based on the numerical measurement outcomes. We challenge the approach where the "exact" value of the observed quantity is compared to the error interval obtained from the measurements under test and we propose a procedure where reference measurements are used as "gold standard". To this purpose, we designed a specific t-test procedure for this purpose, explained here. We also describe and discuss a numerical simulation experiment demonstrating the behaviour of our procedure.

Keywords: measurement; evaluation; Student; t-test; hypothesis testing; interval estimation

1. Introduction

In an experimental context, a number of circumstances require the evaluation of a factor influencing the quality of measurements like the apparatus, its calibration, the context of the measurement set-up and the person(s) performing the measurement. In this paper, we address the question of the optimal use of the measurement outcomes in the evaluation process. Note that we don't exclude the use of supplementary quality criteria in this evaluation process, but we are convinced that the measurement results contain sufficient (complementary) information to justify a more thorough study of their use.

We have chosen as example the evaluation of measurements performed by students in a student lab. A teaching assistant explains to a group of freshmen how a given procedure needs to be performed to measure a specific physical quantity. The students are asked to repeat this procedure a number of times, to calculate from the measurement data a mean measurement value m, as well as an estimate s_m of the standard deviation of m, and to state in their report that they consider the observed physical quantity μ being characterized by an error interval $m \pm s_m$.

The evaluation of the students may include the observation by a teacher or a teaching assistant of the actions by these students during the measurements, as well as the assessment of written reports and/or oral tests. The outcomes of the measurements are a valuable source of information regarding the performances of the students. The measurement data are usually considered to be normally distributed. Although it may be interesting to study the effect on the measurement

*Corresponding author. Email: ehnyssen@etro.vub.ac.be

ISSN: 0266-4763 print/ISSN 1360-0532 online
© 200x Taylor & Francis
DOI: 10.1080/0266476YYxxxxxxxx
http://www.tandfonline.com

procedure of a violation of this simplifying assumption, we will assume its correctness. In the specific pedagogical setting the exact value μ of the measured quantity is assumed being known. The fact whether or not this value is situated in the reported interval, or a scaled version $m \pm as_m$ thereof, is used as a criterion to evaluate the quality of execution of the measurement procedure. We will show in this paper that the described type of approach needs to be challenged from a statistical point of view, and we will propose an alternative for the assessment of the measurement outcomes.

2. Theoretical considerations

Observing a physical quantity by executing a measurement procedure n times is equivalent to drawing a random sample $\{x_1, x_2, \ldots, x_n\}$ from a population of measurement data, distributed about an expectation μ (the actual value of the measured quantity, treated as unknown) with a standard deviation σ. The first steps of error analysis are:

- the calculation of an estimate m for μ, which is the arithmetic sample mean,
- the calculation of an estimate s for σ:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1},$$

- the calculation of an estimate s_m for the standard deviation σ_m of the mean value, which is $s_m = s/\sqrt{n}$ by virtue of the root-n law.

In the measurement evaluation procedure mentioned in Section 1, comparing μ with the scaled error interval $m \pm as_m$ is equivalent to performing a hypothesis test. The sample under test consists of the n measurement outcomes, the test variable is $t_m = (m - \mu)s_m^{-1}$, and the null hypothesis H_0 is $\mathrm{E}\{m\} = \mu$. Indeed, under this hypothesis and the assumption of normality of the measurement outcomes, the distribution of t_m is known to satisfy Student's t-distribution with $n - 1$ degrees of freedom. The test consists of verifying whether $t_m \notin [-a, a]$ and to reject the null hypothesis in this case. This procedure boils down to a verification whether $\mu \notin [m - as_m, m + as_m]$ in which case the validity of the measurements is rejected. The significance level of the test is $1 - P_a$, where $P_a = \mathrm{P}(t_m \in [-a, a])$. One parameter that can be chosen is the sample size n. If this is relatively high (say, $n \geq 30$), s_m can be approximated by σ_m and $t_m \approx (m - \mu)\sigma_m^{-1}$ approximately satisfies a standard normal distribution. The parameter a should be sufficiently high to decrease the significance level. E.g. for high n and $a = 1$, P_1 is only 0.68 which means that correctly performed measurements will only be accepted as such with a probability of 68%! For $n < 30$ this figure is even worse. However, increasing a weakens the test. In many situations, it will be impossible to find a satisfactory trade-off for the choice of a that avoids the wrongful rejection of correct measurements and at the same time yields a criterion to detect bad measurements that has a sufficient sensitivity (the rightful rejection ratio) especially when n is small.

However, our main criticism does not concern the choice of a (e.g. $a = 1$), but the fact that the evaluation criterion is only dependent on the measurements under test. Consequently, independently of the choice of a, an increased value for s_m, which should be interpreted as a decrease in measurement quality, actually leads to a increased acceptance of the measurements.

3. Methods

The main problem is that the parameters μ and σ^2 characterizing correctly acquired measurements are generally unknown. In the present paper, we propose a methodology using the outcomes of a reliable reference measurement as ground truth for a decision on the validity of the measurements under test. In our student evaluation example, this could be realised by letting a skilled teaching assistant or lab technician repeat the measurement process – say, N times. This leads to unbiased estimates m_{R} and s_{R}^2 of the operational parameters μ and σ^2 respectively (provided that the unit(s) of measurement are sufficiently refined for the effect of discretisation to be negligible[1]) – the suffix R stands for "Reference". We will denote by m_{T} and s_{T}^2 the sample mean and variance of the measurements under test (hence, suffix T). Before we describe the measurement evaluation procedure, we formulate the following underlying assumptions:

- correctly acquired measurement data satisfy the normal distribution $\mathrm{N}(\mu, \sigma^2)$,
- the set of reference measurement outcomes are considered as a representative sample from the population of "correctly acquired" measurements, i.e. m_{R} and s_{R}^2, are unbiased estimates of μ and σ^2, respectively.

For comparing the reference measurement data and test measurement data, two criteria are straightforward candidates: the mean value and the sample variance. In this section, we address both of them. In the remainder of the text, we mainly concentrate on the former criterion, because its use is less obvious.

3.1 *Evaluation on the basis of the mean value of the measurement data*

A sound measurement quality assessment procedure only should wrongfully accuse a measurement process of yielding bad outcomes at a specific low rate of, e.g., one to one hundred on average. We designed the formula for the test variable of a hypothesis test that leads to the definition of an acceptance interval for the mean value of the measurement data under evaluation. The design is such that this interval is solely dependent on the reference measurements and on the chosen operational parameters N and n (the number of reference measurements and test measurements respectively). The hypothesis test is based on the following definition of variable t:

$$t = s_{\mathrm{R}}^{-1} \sqrt{\frac{N \cdot n}{N + n}} (m_{\mathrm{R}} - m_{\mathrm{T}}). \tag{1}$$

We can show (see Appendix A) that t satisfies Student's t-distribution with $N - 1$ degrees of freedom under the general assumptions formulated earlier and under the (null) hypothesis that the measurement process under test is correct, implying that the resulting measurements satisfy $\mathrm{N}(\mu, \sigma^2)$.

The acceptance interval can be derived from the relation between $t_{\alpha,N-1}$ and α in

$$\mathrm{P}(-t_{\alpha,N-1} < t < +t_{\alpha,N-1}) = 1 - \alpha. \tag{2}$$

Replacing t by it's expression from Eq. (1), and reformulating the resulting in-

equalities yields

$$\mathrm{P}\left(|m_{\mathrm{T}} - m_{\mathrm{R}}| < t_{\alpha,N-1}\left(\sqrt{\frac{1}{N}+\frac{1}{n}}\right)s_{\mathrm{R}}\right) = 1-\alpha,$$

which defines as acceptance interval for a given α:

$$m_{\mathrm{R}} \pm t_{\alpha,N-1}\left(\sqrt{\frac{1}{N}+\frac{1}{n}}\right)s_{\mathrm{R}}. \tag{3}$$

The critical t-value $t_{\alpha,N-1}$ can be found from the cumulative probability function of t for $N-1$ degrees of freedom - considering that the probability density function is symmetric and consequently Eq. (2) is equivalent to:

$$\mathrm{P}(t < t_{\alpha,N-1}) = 1 - \alpha/2.$$

For a given α, one can find $t_{\alpha,N-1}$ in a table of Student's t distribution – see, e.g., Ref. [5].

The reader may wonder why we don't propose one of the classical t-tests for testing the difference of mean values for independent samples. There exist two variants of these tests: one in which the variances of the data in the two samples are assumed equal and one where this assumption is not required. It is obvious that in general the former model does not hold for the case where test measurements need being compared to reference measurements.

A common formal expression for the second model, is given by:

$$t = \frac{(m_{\mathrm{R}} - m_{\mathrm{T}})}{\sqrt{\frac{s_{\mathrm{R}}^2}{N}+\frac{s_{\mathrm{T}}^2}{n}}} \quad \text{where} \quad df = \left\lceil \frac{\left(\frac{s_{\mathrm{R}}^2}{N}+\frac{s_{\mathrm{T}}^2}{n}\right)^2}{\frac{\left(\frac{s_{\mathrm{R}}^2}{N}\right)^2}{N-1}+\frac{\left(\frac{s_{\mathrm{T}}^2}{n}\right)^2}{n-1}} \right\rceil. \tag{4}$$

The specific equation for df is known as the Welch-Satterthwaite equation.[2] The model is used in studies where the variances of the underlying variable x are allowed to be different in the populations from which the two samples are observed. We already announce here that we dismiss this model as a basis for the evaluation of measurements and refer to sections 4 and 5 for more details about our reasons to do so.

3.2 *Evaluation on the basis of the variance of the measurement data*

Here, we can directly derive the acceptance interval for the variance from a standard F-test. Under the null hypothesis that s_{R} and s_{T} have been calculated from correct measurement data – i.e. two independent samples of data satisfying the same normal distribution $\mathrm{N}(\mu, \sigma^2)$, $F = s_{\mathrm{T}}/s_{\mathrm{R}}$ is distributed according to the F-distribution $\mathrm{F}(n-1, N-1)$.

Assuming that reducing the quality of the measurements will increase the variance of their outcomes, one readily can formulate the acceptance interval as

$$[0\ ,\ s_{\mathrm{R}}\, F_{\alpha,n-1,N-1}],$$

where the critical F-value $F_{\alpha,n-1,N-1}$ is derived from the cumulative probability

function:

$$\mathrm{P}(F < F_{\alpha,n-1,N-1}) = 1 - \alpha.$$

Sometimes, the nature of the measurement process requires considering the possibility that the reduction of the measurement quality either increases or decreases the variance of the outcomes – e.g. when the person performing the measurements systematically tends to round the observed quantities to the same value. In this case the acceptance interval should be:

$$[s_\mathrm{R}\, F^{-1}_{\frac{\alpha}{2},N-1,n-1}\ ,\ s_\mathrm{R}\, F_{\frac{\alpha}{2},n-1,N-1}].$$

4. Numerical Experiments

We performed some numerical experiments to verify in practice the theoretical considerations. Each experiment aimed at estimating the rejection ratio ρ of measurements satisfying one combination of μ_T and σ_T parameter values by repeating the simulation of one measurement experiment a number of times with these specific parameter values. For the estimation of ρ, we calculated the fraction $\hat{\rho}_\mathrm{pt}$ of simulations where the outcomes of the (simulated) measurement experiment are rejected (i.e. a point estimate), as well as an interval estimate of ρ – $[\hat{\rho}_\mathrm{lo}, \hat{\rho}_\mathrm{hi}]$ – at a 95% confidence interval. Evidently, for measurements from a correct measurement process ($\mu_\mathrm{T} = \mu$ and $\sigma_\mathrm{T} = \sigma$), we expect that $\rho(= \mathrm{E}\{\hat{\rho}_\mathrm{pt}\}) = \alpha$. We considered as hypothetical measurement experiment a titration performed by freshmen where the volume of titrant, necessary to neutralize a standardized quantity of acid would be $\mu = 21.35\,\mathrm{cm}^3$. The reference measurements are produced by a laboratory technician performing a titration, repeated $N = 10$ times, with an accuracy characterized by $\sigma = 0.01\,\mathrm{cm}^3$. The mean result obtained by the technician is compared to the mean result obtained by a student who repeats the titration $n = 3$ times and measures according to parameters μ_T and σ_T. In case $\mu_\mathrm{T} = \mu$ and $\sigma_\mathrm{T} = \sigma$, we are dealing with correctly performed measurements. If in that case the student's mean titration volume falls outside the acceptance interval, given by Eq. (3), we are confronted with a wrongful rejection. In order to perform an accurate estimation of the (rightful or wrongful) rejection rate ρ for different combinations of μ_T, σ_T, and α, we simulated 10^6 independent experiments, where each time the student's outcome is tested using the aforementioned acceptance interval. The outcomes of our numerical experiments are summarized in Table 1. In order to compare the criterion based on a scaled error interval to the one presented here, in terms of sensitivity, we repeated some of the experiments with the same parameters and calculating the rejection ratio using the former criterion. The outcomes of these numerical experiments are summarized in Table 2. A selection of the numerical experiments, reported by Table 1 (those for $\mu_\mathrm{T} = 21.35$, $\sigma_\mathrm{T} = 0.01$), have been repeated, but with the criterion based on Eq. (4). The outcomes of these experiments are summarized in Table 3.

5. Discussion

In section 3.1, we propose an method for the evaluation of a measurement process, based on a non-standard Student's t-test, justified theoretically by Appendix A and validated by the numerical experiments described by section 4. Our numerical experiments compare the sensitivity of this methodology with the "classical"

μ_T	σ_T	α	$\hat{\rho}_{pt}$	$\hat{\rho}_{lo}$	$\hat{\rho}_{hi}$
21.35	0.01	0.001	0.00098	0.00092	0.00105
21.35	0.01	0.010	0.01015	0.00995	0.01035
21.35	0.01	0.050	0.04997	0.04954	0.05039
21.37	0.01	0.001	0.13939	0.13872	0.14008
21.37	0.01	0.010	0.46495	0.46397	0.46592
21.37	0.01	0.050	0.77146	0.77064	0.77229
21.35	0.02	0.001	0.02723	0.02691	0.02755
21.35	0.02	0.010	0.10803	0.10742	0.10864
21.35	0.02	0.050	0.24466	0.24382	0.24550

Table 1. Experiments simulating measurements by freshmen compared to a skilled lab technician ($\mu_R = 21.35$, $\sigma_R = 0.01$, $N = 10$, $n = 3$) – results from 10^6 simulations are a point estimate for $\rho - \hat{\rho}_{pt}$ – and an interval estimate for it (confid. level 95%) $[\hat{\rho}_{lo}, \hat{\rho}_{hi}]$ for an acceptance criterion given by Eq. (3).

μ_T	σ_T	α	$\hat{\rho}_{pt}$	$\hat{\rho}_{lo}$	$\hat{\rho}_{hi}$
21.35	0.01	0.010	0.00989	0.00970	0.01009
21.37	0.01	0.010	0.12160	0.12096	0.12224
21.35	0.02	0.010	0.01000	0.00981	0.01020

Table 2. Experiments simulating measurements by freshmen and evaluated on the basis of expression $\mu \in m \pm a s_m$ as acceptance criterion, where a corresponds to $P_a = 1 - \alpha = \mathrm{P}(t_m \in [-a, a])$ as explained in section 2 – results from 10^6 simulations are a point estimate for $\rho - \hat{\rho}_{pt}$ – and an interval estimate $[\hat{\rho}_{lo}, \hat{\rho}_{hi}]$. for it (confid. level 95%).

α	$\hat{\rho}_{pt}$	$\hat{\rho}_{lo}$	$\hat{\rho}_{hi}$
0.001	0.003770	0.003652	0.003892
0.010	0.019548	0.019279	0.019821
0.050	0.065122	0.064640	0.065607

Table 3. Experiments simulating measurements – same parameters as for Table 1 and satisfying H_0, i.e. $\mu_T = 21.35$, $\sigma_T = 0.01$, but with as acceptance/rejection criterion the t-test for the model given by Eq. (4).

approach based on a (scaled) error interval. In the latter approach the value of a is chosen to yield an equivalent criterion as ours in terms of α – the wrongful rejection ratio. The experiments demonstrate that with an appropriate choice of the parameter N, the sensitivity of our approach is much higher than for the (scaled) error interval approach (Tables 1 and 2). Moreover, note also that where a bad measurement procedure only affects σ_T, (last row of Table 2), the sensitivity of the error interval criterion even doesn't exceed α, whereas ours is sensitive both to bias and high σ_T.

The reader could wonder why not using one of the two classical t-tests for independent samples. We already briefly introduced this question at the end of subsection 3.1 and mentioned the hypothesis test of the difference of mean values under the assumption that the variances of the data in the two involved samples are allowed to be different. This problem is known as the Behrens-Fisher problem. In Ref. [3], it is already pointed out that for this problem "There is no completely satisfactory solution known". One very popular solution, found in most text books, is the difference of means test for unequal population variances, using the Welch-Satterthwaite equation – see Eq. (4).[4] One should realize that t in this equation is generally only approximately distributed according to Student's t-distribution. The shortcomings of the technique when both N and n are small, with nevertheless an important discrepancy, are demonstrated by our numerical simulation experiments (Table 3), where under the null hypothesis the wrongful rejection ratio is

systematically significantly larger than the chosen significance level α. Moreover, the sample variances s_R and s_T are treated as equivalent in the calculation of the estimate of the variance of the denominator of the expression for t. In our case, we assume that s_R^2 is an unbiased estimator for σ^2, a parameter that (together with the constants N and n) fully determines the distribution of $(m_\mathrm{R} - m_\mathrm{T})$ under the null hypothesis. It is reasonable to assume that in most cases the variance of the data from badly performed measurements is greater than the variance of the outcomes of correctly performed measurements. In that case, treating the sample variances as equivalent would boil down to weakening the test, especially when considering that normally $n < N$, causing the term in s_T^2 to dominate the term in s_R^2 in the denominator of the expression for t. In the test, proposed here, we decided not to incorporate s_T^2 in the expression for t – see Eq. (1) – and to retain only the "reference" sample standard deviation s_R. An additional advantage is that this approach yields an acceptance interval that only needs to be calculated once for the evaluation of measurement outcomes from several students, since this interval is only dependent on the reference measurements by the laboratory technician. The numerical simulation experiments demonstrate that for correct measurement processes (Table 1 for $\mu_\mathrm{T} = 21.35$ and $\sigma_\mathrm{T} = 0.01$), the wrongful rejection ratio is consistent with the chosen significance level and therefore perfectly controllable. The effect on ρ of a bias in the measurements under evaluation is demonstrated with $\mu_\mathrm{T} = 21.37$ and shows how this ρ (the – this time rightful – rejection rate) increases at the expense of an increasing α. This means for our specific simulated measurements example that if we find such bias of 0.02 sufficiently high for a measurement process to be qualified as bad, and we are satisfied with identifying 77% of the measurement processes affected by such bias, we must accept to wrongfully reject 5% of the correct measurement processes. One may conceive to introduce in the procedure additional information about measurement quality by combining the test on the basis of the mean value of the measurements (subsection 3.1) with the one based on the variance (subsection 3.2), but then one should take care to reduce the α of each test by one half to bring the significance level of the combined test to (at most) α (Bonferroni correction).

References

[1] D.F. Heitjan, *Inference from grouped continuous data: A review*, Stat. Sci. 4 (1989), pp. 164–179.

[2] J. Neter, W. Wasserman, and M.H. Kutner, *Applied linear statistical models : regression, analysis of variance, and experimental designs*, Homewood, Ill. : Irwin (1990).

[3] S.M. Ross, *Introduction to probability and statistics for engineers and scientists*, John Wiley & Sons, Inc (1987).

[4] B.L. Welch, *The significance of the difference between two means when the population variances are unequal*, Biometrika 29 (1938), pp. 350–362.

[5] D. Zwillinger (ed.), *CRC – Standard mathematical tables and formulæ*, 31st ed., Chapman & Hall/CRC (2003).

Appendix A. Student evaluation based on the mean measurement value – theoretical background

The goal of this appendix is to demonstrate that the expression for t in Eq. (1) satisfies Student's t-distribution with $N - 1$ degrees of freedom.

Let us start from the formal definition of Student's t-distributed variable, con-

sisting of the following elements:[5]

(D1) If the random variable X is normally distributed with mean 0 and variance σ^2, and ...
(D2) ...if Y^2/σ^2 has a χ^2 distribution with ν degrees of freedom, and ...
(D3) ...if X and Y are independent, ...
(D4) ...then $t = \frac{X\sqrt{\nu}}{Y}$ is distributed as a t-distribution with ν degrees of freedom.

As linear combination of normally distributed terms m_{R} and m_{T}, expression $(m_{\mathrm{R}} - m_{\mathrm{T}})$ satisfies a normal distribution. Since m_{R} and m_{T} have the same expectation μ, the expectation of $(m_{\mathrm{R}} - m_{\mathrm{T}})$ is zero. Also, we are dealing with the mean values of independent sets of measurement data. Therefore m_{R} and m_{T} are statistically independent, resulting in the variance of $(m_{\mathrm{R}} - m_{\mathrm{T}})$ being the sum of the variances of m_{R} and m_{T}. Formally:

$$(m_{\mathrm{R}} - m_{\mathrm{T}}) : \mathrm{N}\left(0, \sigma^2\left(\frac{1}{N} + \frac{1}{n}\right)\right).$$

This allows us to introduce a variable X and to equate it to an expression that satisfies element (D1) of the definition of the t-distribution, formulated earlier:

$$X = \sqrt{\frac{N \cdot n}{N + n}}(m_{\mathrm{R}} - m_{\mathrm{T}}) \quad : \quad \mathrm{N}(0, \sigma^2). \tag{A1}$$

Let us now introduce a variable

$$Y = \sqrt{\nu}\, s_{\mathrm{R}} \quad \text{with} \quad \nu = N - 1. \tag{A2}$$

From this definition, and the fundamental properties of the sample variance of a normally distributed variable follows that Y^2/σ^2 has a χ^2 distribution with ν degrees of freedom, which satisfies element (D2) of the aforementioned definition.

Finally note that X and Y are independent – definition element (D3) – since on the one hand m_{R} and s_{R} are mutually independent as mean and variance of data of the same sample and on the other hand, m_{T} and s_{R} are independent as statistics of two independent sets of sample data.

This means that $t = \frac{X\sqrt{\nu}}{Y}$ satisfies the t-distribution according to definition element (D4), but substituting X, Y and ν in the latter expression for t, using Eqs. (A1) and (A2) yields exactly Eq. (1).